ANALYTICS *in a* BIG DATA WORLD

The ESSENTIAL GUIDE *to* DATA SCIENCE *and its* APPLICATIONS

BART BAESENS

WILEY

Analytics in a Big Data World

The Essential Guide to Data Science and Its Applications

Bart Baesens

WILEY

Cover image: ©iStockphoto/vlastos
Cover design: Wiley

Published by John Wiley & Sons, Inc., Hoboken, New Jersey.
Published simultaneously in Canada.

Library of Congress Cataloging-in-Publication Data:
Baesens, Bart.
Analytics in a big data world : the essential guide to data science and its applications / Bart Baesens.
1 online resource. — (Wiley & SAS business series)
Description based on print version record and CIP data provided by publisher; resource not viewed.
ISBN 978-1-118-89271-8 (ebk); ISBN 978-1-118-89274-9 (ebk); ISBN 978-1-118-89270-1 (cloth) 1. Big data. 2. Management—Statistical methods. 3. Management—Data processing. 4. Decision making—Data processing. I. Title.
HD30.215
658.4'038 dc23

2014004728

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

Analytics in a Big Data World

Wiley & SAS Business Series

The Wiley & SAS Business Series presents books that help senior-level managers with their critical management decisions.

Titles in the Wiley & SAS Business Series include:

Activity-Based Management for Financial Institutions: Driving Bottom-Line Results by Brent Bahnub

Bank Fraud: Using Technology to Combat Losses by Revathi Subramanian

Big Data Analytics: Turning Big Data into Big Money by Frank Ohlhorst

Branded! How Retailers Engage Consumers with Social Media and Mobility by Bernie Brennan and Lori Schafer

Business Analytics for Customer Intelligence by Gert Laursen

Business Analytics for Managers: Taking Business Intelligence beyond Reporting by Gert Laursen and Jesper Thorlund

The Business Forecasting Deal: Exposing Bad Practices and Providing Practical Solutions by Michael Gilliland

Business Intelligence Applied: Implementing an Effective Information and Communications Technology Infrastructure by Michael Gendron

Business Intelligence in the Cloud: Strategic Implementation Guide by Michael S. Gendron

Business Intelligence Success Factors: Tools for Aligning Your Business in the Global Economy by Olivia Parr Rud

CIO Best Practices: Enabling Strategic Value with Information Technology, second edition by Joe Stenzel

Connecting Organizational Silos: Taking Knowledge Flow Management to the Next Level with Social Media by Frank Leistner

Credit Risk Assessment: The New Lending System for Borrowers, Lenders, and Investors by Clark Abrahams and Mingyuan Zhang

Credit Risk Scorecards: Developing and Implementing Intelligent Credit Scoring by Naeem Siddiqi

The Data Asset: How Smart Companies Govern Their Data for Business Success by Tony Fisher

Delivering Business Analytics: Practical Guidelines for Best Practice by Evan Stubbs

Demand-Driven Forecasting: A Structured Approach to Forecasting, Second Edition by Charles Chase

Demand-Driven Inventory Optimization and Replenishment: Creating a More Efficient Supply Chain by Robert A. Davis

The Executive's Guide to Enterprise Social Media Strategy: How Social Networks Are Radically Transforming Your Business by David Thomas and Mike Barlow

Economic and Business Forecasting: Analyzing and Interpreting Econometric Results by John Silvia, Azhar Iqbal, Kaylyn Swankoski, Sarah Watt, and Sam Bullard

Executive's Guide to Solvency II by David Buckham, Jason Wahl, and Stuart Rose

Fair Lending Compliance: Intelligence and Implications for Credit Risk Management by Clark R. Abrahams and Mingyuan Zhang

Foreign Currency Financial Reporting from Euros to Yen to Yuan: A Guide to Fundamental Concepts and Practical Applications by Robert Rowan

Health Analytics: Gaining the Insights to Transform Health Care by Jason Burke

Heuristics in Analytics: A Practical Perspective of What Influences Our Analytical World by Carlos Andre Reis Pinheiro and Fiona McNeill

Human Capital Analytics: How to Harness the Potential of Your Organization's Greatest Asset by Gene Pease, Boyce Byerly, and Jac Fitz-enz

Implement, Improve and Expand Your Statewide Longitudinal Data System: Creating a Culture of Data in Education by Jamie McQuiggan and Armistead Sapp

Information Revolution: Using the Information Evolution Model to Grow Your Business by Jim Davis, Gloria J. Miller, and Allan Russell

Killer Analytics: Top 20 Metrics Missing from Your Balance Sheet by Mark Brown

Manufacturing Best Practices: Optimizing Productivity and Product Quality by Bobby Hull

Marketing Automation: Practical Steps to More Effective Direct Marketing by Jeff LeSueur

Mastering Organizational Knowledge Flow: How to Make Knowledge Sharing Work by Frank Leistner

The New Know: Innovation Powered by Analytics by Thornton May

Performance Management: Integrating Strategy Execution, Methodologies, Risk, and Analytics by Gary Cokins

Predictive Business Analytics: Forward-Looking Capabilities to Improve Business Performance by Lawrence Maisel and Gary Cokins

Retail Analytics: The Secret Weapon by Emmett Cox

Social Network Analysis in Telecommunications by Carlos Andre Reis Pinheiro

Statistical Thinking: Improving Business Performance, second edition by Roger W. Hoerl and Ronald D. Snee

Taming the Big Data Tidal Wave: Finding Opportunities in Huge Data Streams with Advanced Analytics by Bill Franks

Too Big to Ignore: The Business Case for Big Data by Phil Simon

The Value of Business Analytics: Identifying the Path to Profitability by Evan Stubbs

Visual Six Sigma: Making Data Analysis Lean by Ian Cox, Marie A. Gaudard, Philip J. Ramsey, Mia L. Stephens, and Leo Wright

Win with Advanced Business Analytics: Creating Business Value from Your Data by Jean Paul Isson and Jesse Harriott

For more information on any of the above titles, please visit www.wiley.com.

To my wonderful wife, Katrien, and my kids,
Ann-Sophie, Victor, and Hannelore.
To my parents and parents-in-law.

Preface

Companies are being flooded with tsunamis of data collected in a multichannel business environment, leaving an untapped potential for analytics to better understand, manage, and strategically exploit the complex dynamics of customer behavior. In this book, we will discuss how analytics can be used to create strategic leverage and identify new business opportunities.

The focus of this book is not on the mathematics or theory, but on the practical application. Formulas and equations will only be included when absolutely needed from a practitioner's perspective. It is also not our aim to provide exhaustive coverage of all analytical techniques previously developed, but rather to cover the ones that really provide added value in a business setting.

The book is written in a condensed, focused way because it is targeted at the business professional. A reader's prerequisite knowledge should consist of some basic exposure to descriptive statistics (e.g., mean, standard deviation, correlation, confidence intervals, hypothesis testing), data handling (using, for example, Microsoft Excel, SQL, etc.), and data visualization (e.g., bar plots, pie charts, histograms, scatter plots). Throughout the book, many examples of real-life case studies will be included in areas such as risk management, fraud detection, customer relationship management, web analytics, and so forth. The author will also integrate both his research and consulting experience throughout the various chapters. The book is aimed at senior data analysts, consultants, analytics practitioners, and PhD researchers starting to explore the field.

Chapter 1 discusses big data and analytics. It starts with some example application areas, followed by an overview of the analytics process model and job profiles involved, and concludes by discussing key analytic model requirements. Chapter 2 provides an overview of

data collection, sampling, and preprocessing. Data is the key ingredient to any analytical exercise, hence the importance of this chapter. It discusses sampling, types of data elements, visual data exploration and exploratory statistical analysis, missing values, outlier detection and treatment, standardizing data, categorization, weights of evidence coding, variable selection, and segmentation. Chapter 3 discusses predictive analytics. It starts with an overview of the target definition and then continues to discuss various analytics techniques such as linear regression, logistic regression, decision trees, neural networks, support vector machines, and ensemble methods (bagging, boosting, random forests). In addition, multiclass classification techniques are covered, such as multiclass logistic regression, multiclass decision trees, multiclass neural networks, and multiclass support vector machines. The chapter concludes by discussing the evaluation of predictive models. Chapter 4 covers descriptive analytics. First, association rules are discussed that aim at discovering intratransaction patterns. This is followed by a section on sequence rules that aim at discovering intertransaction patterns. Segmentation techniques are also covered. Chapter 5 introduces survival analysis. The chapter starts by introducing some key survival analysis measurements. This is followed by a discussion of Kaplan Meier analysis, parametric survival analysis, and proportional hazards regression. The chapter concludes by discussing various extensions and evaluation of survival analysis models. Chapter 6 covers social network analytics. The chapter starts by discussing example social network applications. Next, social network definitions and metrics are given. This is followed by a discussion on social network learning. The relational neighbor classifier and its probabilistic variant together with relational logistic regression are covered next. The chapter ends by discussing egonets and bigraphs. Chapter 7 provides an overview of key activities to be considered when putting analytics to work. It starts with a recapitulation of the analytic model requirements and then continues with a discussion of backtesting, benchmarking, data quality, software, privacy, model design and documentation, and corporate governance. Chapter 8 concludes the book by discussing various example applications such as credit risk modeling, fraud detection, net lift response modeling, churn prediction, recommender systems, web analytics, social media analytics, and business process analytics.

Acknowledgments

I would like to acknowledge all my colleagues who contributed to this text: Seppe vanden Broucke, Alex Seret, Thomas Verbraken, Aimée Backiel, Véronique Van Vlasselaer, Helen Moges, and Barbara Dergent.

Analytics in a Big Data World

Contents

CHAPTER 1

Big Data and Analytics

Data are everywhere. IBM projects that every day we generate 2.5 quintillion bytes of data.[1] In relative terms, this means 90 percent of the data in the world has been created in the last two years. Gartner projects that by 2015, 85 percent of Fortune 500 organizations will be unable to exploit big data for competitive advantage and about 4.4 million jobs will be created around big data.[2] Although these estimates should not be interpreted in an absolute sense, they are a strong indication of the ubiquity of big data and the strong need for analytical skills and resources because, as the data piles up, managing and analyzing these data resources in the most optimal way become critical success factors in creating competitive advantage and strategic leverage.

Figure 1.1 shows the results of a KDnuggets[3] poll conducted during April 2013 about the largest data sets analyzed. The total number of respondents was 322 and the numbers per category are indicated between brackets. The median was estimated to be in the 40 to 50 gigabyte (GB) range, which was about double the median answer for a similar poll run in 2012 (20 to 40 GB). This clearly shows the quick increase in size of data that analysts are working on. A further regional breakdown of the poll showed that U.S. data miners lead other regions in big data, with about 28% of them working with terabyte (TB) size databases.

A main obstacle to fully harnessing the power of big data using analytics is the lack of skilled resources and "data scientist" talent required to

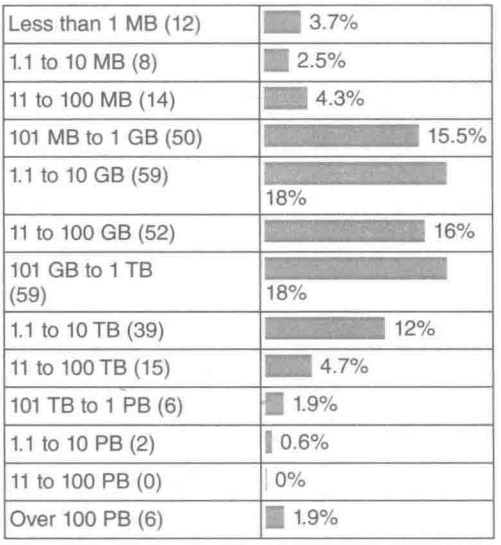

Figure 1.1 Results from a KDnuggets Poll about Largest Data Sets Analyzed
Source: www.kdnuggets.com/polls/2013/largest-dataset-analyzed-data-mined-2013.html.

exploit big data. In another poll ran by KDnuggets in July 2013, a strong need emerged for analytics/big data/data mining/data science education.[4] It is the purpose of this book to try and fill this gap by providing a concise and focused overview of analytics for the business practitioner.

EXAMPLE APPLICATIONS

Analytics is everywhere and strongly embedded into our daily lives. As I am writing this part, I was the subject of various analytical models today. When I checked my physical mailbox this morning, I found a catalogue sent to me most probably as a result of a response modeling analytical exercise that indicated that, given my characteristics and previous purchase behavior, I am likely to buy one or more products from it. Today, I was the subject of a behavioral scoring model of my financial institution. This is a model that will look at, among other things, my checking account balance from the past 12 months and my credit payments during that period, together with other kinds of information available to my bank, to predict whether I will default on my loan during the next year. My bank needs to know this for provisioning purposes. Also today, my telephone services provider analyzed my calling behavior